



e-ISSN: 2319-8753 | p-ISSN: 2320-6710

IJRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 5, May 2021

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.512

9940 572 462

6381 907 438

ijirset@gmail.com

www.ijirset.com

Fake News Detection using Machine Learning

Uzeena Mhaskar¹, Prerana Kadam², Pragati Ghadashi³, Swati Powar⁴

U.G. Student, Department of Information Technology, FAMT Engineering College, Ratnagiri, Maharashtra, India^{1,2,3}

Associate Professor, Department of Information Technology, FAMT Engineering College, Ratnagiri,
Maharashtra, India⁴

ABSTRACT: The rapid growth of social networking platforms has not only yielded a enormous increase in data accessibility however has accelerated the spread of fake news. Fake news can take advantage of multimedia content to mislead readers and gets spread, which may cause negative effects or maybe manipulate the general public events. The goal of the fake news websites is to affect public opinion on certain matters. Our aim is to search out a reliable and correct model that classifies a given article and check whether or not its fake or real. This paper helps to notice the fake news and its accuracy using machine learning algorithms.

KEYWORDS: Fake News Detection, Logistic regression, Python, Naïve Bayes, Data Cleaning, Data Pre-processing.

I. INTRODUCTION

Modern life has become quite appropriate and therefore the individuals of the planet ought to give thanks the huge contribution of the net technology for transmission and knowledge sharing. There is no doubt that internet has made our lives easier and access to excessive information. Nowadays, there are increasingly more people consuming news through social media, which may provide timely and comprehensive multimedia data on the events taking place all over the world. Social media taking part in an enormous role in influencing the society and because it is common, some people try to take advantage of it. Compared with ancient text news, the news with pictures and videos will provide a better storytelling and attract more attention from readers and thus fake news accumulates a good deal of attention over the internet particularly on social media. It contains twisted or maybe cast pictures, to mislead the readers and find speedy dissemination. The dissemination of fake news will cause large-scale negative effects and typically will have an effect on or maybe manipulate vital public events. Individuals get deceived and do not consider before circulating such mis-informative items to the far end of the world. This sort of reports vanishes but not without doing the harm it intended to cause. Their primary purpose is to govern the data that can make public believe it. There are numerous examples of such websites all over the world. The social media sites that play a serious role in supply of fake news include Facebook, Twitter, WhatsApp, etc that have an effect on the minds of the individuals.

According to study many Scientist believe that counterfeited news issue could also be addressed by means of machine learning and artificial intelligence. This is because the result of the synthetic intelligence is currently changing into very talked-about and lots of devices are offered to ascertain it part. Recently computer science algorithms have begun to enhance work on numerous classification issues (image recognition, voice detection and so on) as a result of hardware is cheaper and larger datasets are offered. Various models are used to detect and provide an accuracy range that comprises of clustering, deception modelling, predictive modelling and various algorithms like Bayes, Decision trees, SVM, Logistic Regression, Random Forest, etc. An algorithm has been explored that can distinguish the difference between the false and true news with accuracy on the basis of certain criteria which are spelling mistakes, jumbled sentences, punctuation errors etc. The aim of this paper is to extend the accuracy of detecting fake news more than the present results.



II. LITERATURE SURVEY

Social media is not just used to share opinion but also news, as well as rumours and fake news. As there's no strict check on tweets, individuals tend to post false news for his or her own profit. Heaps of individuals retweet it and it spreads like a pandemic.

Many of us are in the loop for trying to reduce the harmful effects it's inflicting. [1] Janze et al. trained a range of machine learning classifiers suitable for binary classification drawback i.e. Logistic regression, Support vector machines, Random forest. They divided the dataset into equally sized folds containing same amount of fake and real observations. They obtained an accuracy of nearly 80%. [2] Gupta et al. conferred a path breaking work by characterizing and identifying fake pictures on twitter during Hurricane Sandy. They said that 86% of the tweets containing fake pictures were retweets. There were very few instances of users posting original tweets with fake images. URL shorteners were worn to spread the fake news. [3] In follow-up work, Aditi Gupta wrote on credibility ranking of tweets during huge impact events as 30% of the tweets posted contained situational data while 14% was spam. They used linear logistic regression, SVM ranking algorithm and supervised learning for their work.

The aim of this review is to assemble the assorted sequence of related work done on the area of fake news detection over the social network. This review isn't specific to certain sectors i.e. the health sector or the business enterprise sector however rather consider on all elements that contribute to individuals sharing false information. Thereby, we tend to proceed the survey from numerous datasets with the intention of detecting the possible reliability of the worldly knowledge.

III. METHODOLOGY

I. Dataset

Getting the correct information implies assembling or distinguishing the information that relates with the results that has to be predicted. Selecting the correct dataset for Machine learning is incredibly vital to make the AI model functional with right approach. Though selecting the right quality and amount of data is challenging task however there are few rules that has to be followed for machine learning on big data.

In this project, the dataset is being taken from kaggle.com. The size of the dataset is 6336. There are 6336 rows and three columns. Currently it's to be seen the accuracy that the rule will offer.

II. Pre-Processing

Data pre-processing is a method that is used to change over the unrefined information into an ideal informational index. At the end of the day, at whatever point the data is assembled from various sources it's gathered in associate unrefined organization that is not possible for the examination. Pre-processing is important for accomplishing higher outcomes from the applied model in Machine Learning project the configuration of the data should be in a permitted way. This helps reduce the size of the actual data by removing the irrelevant information that exists within the information.

III. Train and Test Splitting

Machine learning frequently works with two informational collections: training and testing. Each of the two ought to arbitrarily test a bigger assortment of data. The principal set that is getting used is the training set, the biggest of the two. Running a training set through a machine learning system shows internet to weigh diverse highlights, changing them coefficients as per their probability of limiting blunders within the outcomes. Those coefficients, otherwise referred to as parameters, are contained in tensors and together they are called the model, since it encodes a model of the data on which it is being trained. They are the most vital takeaways that is being acquired from making ready a machine learning system. The second set is the test set. It works as a seal of endorsement, and is not used till the end. Once it is being ready and information is set, the neural internet may be tested against this last arbitrary examination.

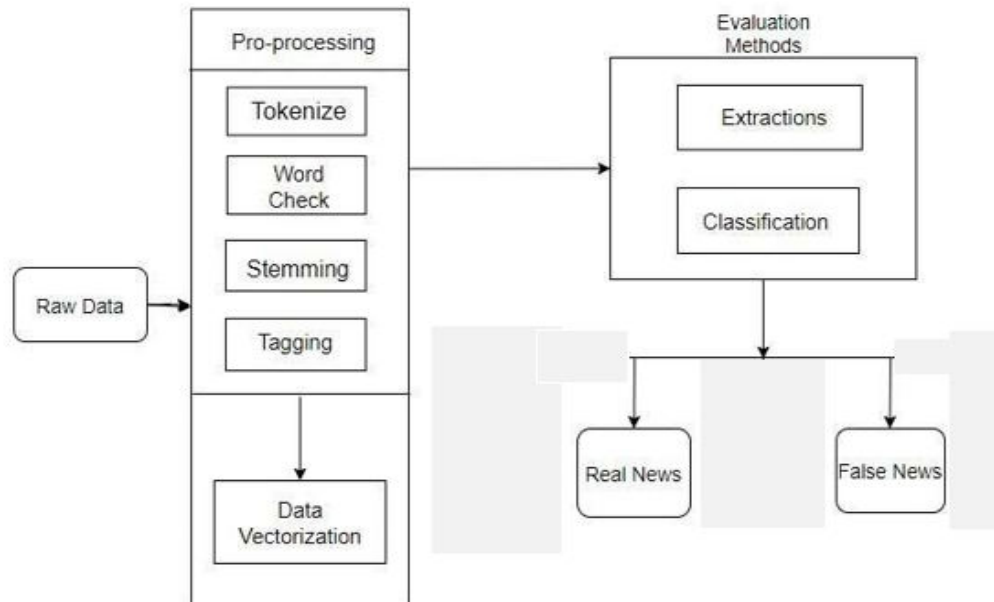


Fig 1. System Representation

IV. EVALUATION MODELS

I. Logistic Regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of categories. Unlike linear regression which outputs continuous number values, logistic regression changes its yield utilizing the calculated sigmoid capability to restore a likelihood esteem which might then be able to be mapped to a minimum of 2 distinct categories. The LR model uses gradient descent to converge onto the optimal set of weights (θ) for the training set.

II. Support Vector Machine

A Support Vector Machine (SVM) is a supervised machine learning algorithm that is used for both classification and regression. It is used for the data that is not frequently distributed and have unknown distributions. The SVM provides a very useful technique among it known as kernel and by the application of associated kernel function we can solve any complex problem.

III. Naïve Bayes Classifier

Naïve Bayes is a machine learning algorithm use to solve classification issues. It is one of the simplest yet powerful ML algorithms in use and finds applications in several industries. Naive Bayes classifiers are a family of straightforward "probabilistic classifiers" based on applying Bayes' theorem with powerful (naive) independent assumptions between the features. Naive Bayes classifiers are highly scalable, requiring variety of parameters linear within the number of variables (features/predictors) during a learning problem. The formula for naive Bayes classifier is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)}$$



V. IMPLEMENTATION

First we import all the required libraries.

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pylab as pl
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.linear_model import PassiveAggressiveClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import TfidfVectorizer
import pickle
```

Fig 2.Importing Libraries

Then we load the dataset using the function pd.read_csv.Here we have the headers of the dataset.

```
In [2]: news = pd.read_csv('news.csv')
X = news['text']
y = news['label']
```

```
In [3]: news.head()
```

Out[3]:

	Unnamed: 0		title	text	label
0	8476		You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...		Google Pinterest Digg LinkedIn Reddit Stumbleu...	FAKE
2	3608		Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...		— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters		It's primary day in New York and front-runners...	REAL

Fig 3.Loading the Dataset

Here we have the dataset of 6335 rows and 4 columns.

```
In [4]: df=pd.read_csv('news.csv')
print(df)
```

```
6305 Print \n[Ed. - Now teaching the gospel of raci... FAKE
6306 Sound too "strange" to be true? We have proof!... FAKE
6307 US abstains from UN vote calling for end to Cu... FAKE
6308 Tuesday 1 November 2016 by Formelia Alberthine... FAKE
6309 When Susan E. Rice took over as President Obam... REAL
6310 (CNN) ISIS has claimed responsibility for the ... REAL
6311 It was inevitable that liberals would end up b... REAL
6312 By Kalee Brown\nWhile I was at university, man... FAKE
6313 Email \nDonald Trump is again riling up his vo... FAKE
6314 BREAKING: Trump Jumps in FL, Takes 4 Point Lea... FAKE
6315 Police in Charleston, S.C., say a man they sus... REAL
6316 Silver of FiveThirtyEight.com has laid out fou... REAL
6317 This post was originally published on this sit... FAKE
6318 BREAKING : Hillary Campaign Manager Deletes hi... FAKE
6319 Ted Cruz took first prize in the Iowa caucuses... REAL
6320 Posted on November 4, 2016 by Charles Hugh Smi... FAKE
6321 Charlie Baker , Massachusetts (2015-present)[3... FAKE
6322 vladimir putin , Valdai , sochi , RBTH Daily R... FAKE
6323 ROME - U.S. Democratic presidential candidate... REAL
6324 Most conservatives who oppose marriage equalit... REAL
6325 Written by Peter Van Buren venerable New Yor... FAKE
6326 DOJ COMPLAINT: Comey Under Fire Over Partisan ... FAKE
6327 The freshman senator from Georgia quoted scrip... REAL
6328 FAKE
6329 Julian Assange has claimed the Hillary Clinton... FAKE
6330 The State Department told the Republican Natio... REAL
6331 The 'P' in PBS Should Stand for 'Plutocratic' ... FAKE
6332 Anti-Trump Protesters Are Tools of the Oligar... FAKE
6333 ADDIS ABABA, Ethiopia -President Obama convene... REAL
6334 Jeb Bush Is Suddenly Attacking Trump. Here's W... REAL

[6335 rows x 4 columns]
```

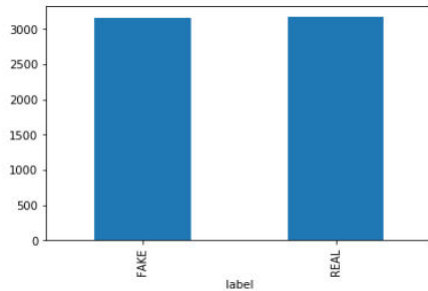
Fig4. Total number of rows and columns in the dataset.



Here is the graph for our project where it shows how many articles are fake and real.

```
In [5]: print(news.groupby(['label'])['text'].count())
news.groupby(['label'])['text'].count().plot(kind="bar")
plt.show()
```

```
label
FAKE    3164
REAL    3171
Name: text, dtype: int64
```



```
In [6]: news.shape
```

```
Out[6]: (6335, 4)
```

Fig 5.Bar plot for total no. of fake and real news articles.

We then perform data pre-processing by cleaning the data by checking the null values in dataset using function `isnull.any()`.

```
In [7]: news.isnull().any()
```

```
Out[7]: Unnamed: 0    False
title           False
text            False
label           False
dtype: bool
```

```
In [8]: news.isnull().sum()
```

```
Out[8]: Unnamed: 0    0
title           0
text            0
label           0
dtype: int64
```

Fig 6.Data pre-processing

We divide the data into train and tests sets which is in ratio 80:20 where 80% is train set and 20% is test set. Now we implement Tfidf vectorizer which is Term Frequency Inverse Document Frequency which finds the importance of keyword in web page. It also removes the stop words. So these are the words that commonly appear in web page such as and, like, and other punctuation.

```
In [9]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
In [10]: tfidf_vectorizer=TfidfVectorizer(stop_words='english',max_df=0.7)
tfidf_train=tfidf_vectorizer.fit_transform(X_train)
tfidf_test=tfidf_vectorizer.transform(X_test)
```

Fig 7.Dividing Data into trained and test sets

After that we implement the passive aggressive classifier which is an online learning algorithm that reminds passive for classification and turns aggressive for mis-calculation. It is used to calculate the accuracy. So here we get accuracy of



93.76%.

```
In [11]: pac=PassiveAggressiveClassifier(max_iter=50)
pac.fit(tfidf_train,y_train)
y_pred=pac.predict(tfidf_test)
score=accuracy_score(y_test,y_pred)
print(f'Accuracy: {round(score*100,2)}%')

Accuracy: 93.76%
```

Fig 8.Implementing Passive Aggressive Classifier

After these we implement the pipeline utility function which is used to train data, transform to test data. So we remove stop words first and then apply MultinomialNB. After implementing pipeline utility function we get an accuracy score of 0.836621.

```
In [12]: pipeline = Pipeline([('tfidf', TfidfVectorizer(stop_words='english')),
('nbmodel', MultinomialNB())])

In [13]: pipeline.fit(X_train, y_train)
pred = pipeline.predict(X_test)

In [14]: score=pipeline.score(X_test,y_test)
print('accuracy',score)

accuracy 0.8366219415943172
```

Fig 9.Pipeline Utility function with multinomialNB

Now we print the performance evaluation table and also the confusion matrix.

```
In [15]: print(classification_report(y_test, pred))
```

	precision	recall	f1-score	support
FAKE	0.97	0.70	0.81	644
REAL	0.76	0.98	0.85	623
accuracy			0.84	1267
macro avg	0.86	0.84	0.83	1267
weighted avg	0.87	0.84	0.83	1267

```
In [16]: print(confusion_matrix(y_test, pred))
```

```
[[452 192]
 [ 15 608]]
```

```
In [17]: with open('model.pickle', 'wb') as handle:
pickle.dump(pipeline, handle, protocol=pickle.HIGHEST_PROTOCOL)
```

```
In [ ]:
```

Fig 10. Confusion Matrix

After these we then go to synder and execute app.py code here. First we import all the required libraries and then load the flask app. Flask is web application framework and then we assign the model variable.

```
#Importing the Libraries
import numpy as np
from flask import Flask, request, render_template
from flask_cors import CORS
import os
from sklearn.externals import joblib
import pickle
import flask
import os
import newspaper
from newspaper import Article
import urllib

#Loading Flask and assigning the model variable
app = Flask(__name__)
CORS(app)
app=flask.Flask(__name__, template_folder='templates')

with open('model.pickle', 'rb') as handle:
    model = pickle.load(handle)

@app.route('/')
def main():
    return render_template('main.html')
```

Fig 11.Loading flask app

In these part of code we receive the URL from user as input and use web scraping to extract the news content.

```
#Receiving the input url from the user and using Web Scrapping to extract the news content
@app.route('/predict', methods=['GET', 'POST'])
def predict():
    url =request.get_data(as_text=True)[5:]
    url = urllib.parse.unquote(url)
    article = Article(str(url))
    article.download()
    article.parse()
    article.nlp()
    news = article.summary
```

Fig 12.URL for website

Where web scraping is process of extracting data from websites.After passing news article to model we implement “pred” part of code to check whether the given news article is fake or real.

```
#Passing the news article to the model and returning whether it is Fake or Real
pred = model.predict([news])
return render_template('main.html', prediction_text='The news is "{}".format(pred[0]))

if __name__=="__main__":
    port=int(os.environ.get('PORT',5000))
    app.run(port=port,debug=True,use_reloader=False)
```

Fig 13.Web scraping

Here we have use index.html file for the UI of our project.For these User Interface we use style.css file to describe the presentation of website.



```
@import url(https://fonts.googleapis.com/css?family=Open+Sans);
.btn { display: inline-block; *display: inline; *zoom: 1; padding: 4px 10px 4px; margin-bottom: 0;
.btn:hover, .btn:active, .btn.active, .btn.disabled, .btn[disabled] { background-color: #e6e6e6; }
.btn-large { padding: 9px 14px; font-size: 15px; line-height: normal; -webkit-border-radius: 5px;
.btn:hover { color: #333333; text-decoration: none; background-color: #e6e6e6; background-position:
.btn-primary, .btn-primary:hover { text-shadow: 0 -1px 0 rgba(0, 0, 0, 0.25); color: #ffffff; }
.btn-primary.active { color: rgba(255, 255, 255, 0.75); }
.btn-primary { background-color: #4a77d4; background-image: -moz-linear-gradient(top, #6eb6de, #4a77d4);
.btn-primary:hover, .btn-primary:active, .btn-primary.active, .btn-primary.disabled, .btn-primary[disabled] {
.btn-block { width: 100%; display:block; }

* { -webkit-box-sizing:border-box; -moz-box-sizing:border-box; -ms-box-sizing:border-box; -o-box-sizing:border-box; box-sizing:border-box; }

html { width: 100%; height:100%; overflow:hidden; }

body {
width: 100%;
height:100%;
font-family: 'Open Sans', sans-serif;
background: #092756;
color: #fff;
font-size: 18px;
text-align:center;
letter-spacing:1.2px;
background: -moz-radial-gradient(0% 100%, ellipse cover, rgba(104,128,138,.4) 10%,rgba(138,114,76,0.4) 40%,
background: -webkit-radial-gradient(0% 100%, ellipse cover, rgba(104,128,138,.4) 10%,rgba(138,114,76,0.4) 40%,
background: -o-radial-gradient(0% 100%, ellipse cover, rgba(104,128,138,.4) 10%,rgba(138,114,76,0.4) 40%,
background: -ms-radial-gradient(0% 100%, ellipse cover, rgba(104,128,138,.4) 10%,rgba(138,114,76,0.4) 40%,
background: -webkit-radial-gradient(0% 100%, ellipse cover, rgba(104,128,138,.4) 10%,rgba(138,114,76,0.4) 40%,
filter: progid:DXImageTransform.Microsoft.gradient( startColorstr='#3E1D6D', endColorstr='#092756', gradientType=radial);
}
.login {
```

Fig 14.Styling for UI

We then compile these project in Anaconda prompt. We now get a URL for our website.

```
Anaconda Prompt (Anaconda3) - python app.py
(base) C:\Users\UZEENA>cd Fake_news_detection
(base) C:\Users\UZEENA\Fake_news_detection>python app.py
C:\Users\UZEENA\Anaconda3\lib\site-packages\sklearn\externals\joblib\_init__.py:15: DeprecationWarning: sklearn.externals.joblib is deprecated in 0.21 and will be removed in 0.23. Please import this functionality directly from joblib, which can be installed with: pip install joblib. If this warning is raised when loading pickled models, you may need to re-serialize those models with scikit-learn 0.21+.
  warnings.warn(msg, category=DeprecationWarning)
C:\Users\UZEENA\Anaconda3\lib\site-packages\sklearn\base.py:306: UserWarning: Trying to unpickle estimator TfidfTransformer from version 0.21.3 when using version 0.21.2. This might lead to breaking code or invalid results. Use at your own risk.
  UserWarning()
C:\Users\UZEENA\Anaconda3\lib\site-packages\sklearn\base.py:306: UserWarning: Trying to unpickle estimator TfidfVectorizer from version 0.21.3 when using version 0.21.2. This might lead to breaking code or invalid results. Use at your own risk.
  UserWarning()
C:\Users\UZEENA\Anaconda3\lib\site-packages\sklearn\base.py:306: UserWarning: Trying to unpickle estimator MultinomialNB from version 0.21.3 when using version 0.21.2. This might lead to breaking code or invalid results. Use at your own risk.
  UserWarning()
C:\Users\UZEENA\Anaconda3\lib\site-packages\sklearn\base.py:306: UserWarning: Trying to unpickle estimator Pipeline from version 0.21.3 when using version 0.21.2. This might lead to breaking code or invalid results. Use at your own risk.
  UserWarning()
* Serving Flask app "app" (lazy loading)
* Environment: production
  WARNING: This is a development server. Do not use it in a production deployment.
  Use a production WSGI server instead.
* Debug mode: on
* Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

Fig 15.Anaconda Prompt to compile project



IV. RESULTS

So these is the UI of our project.

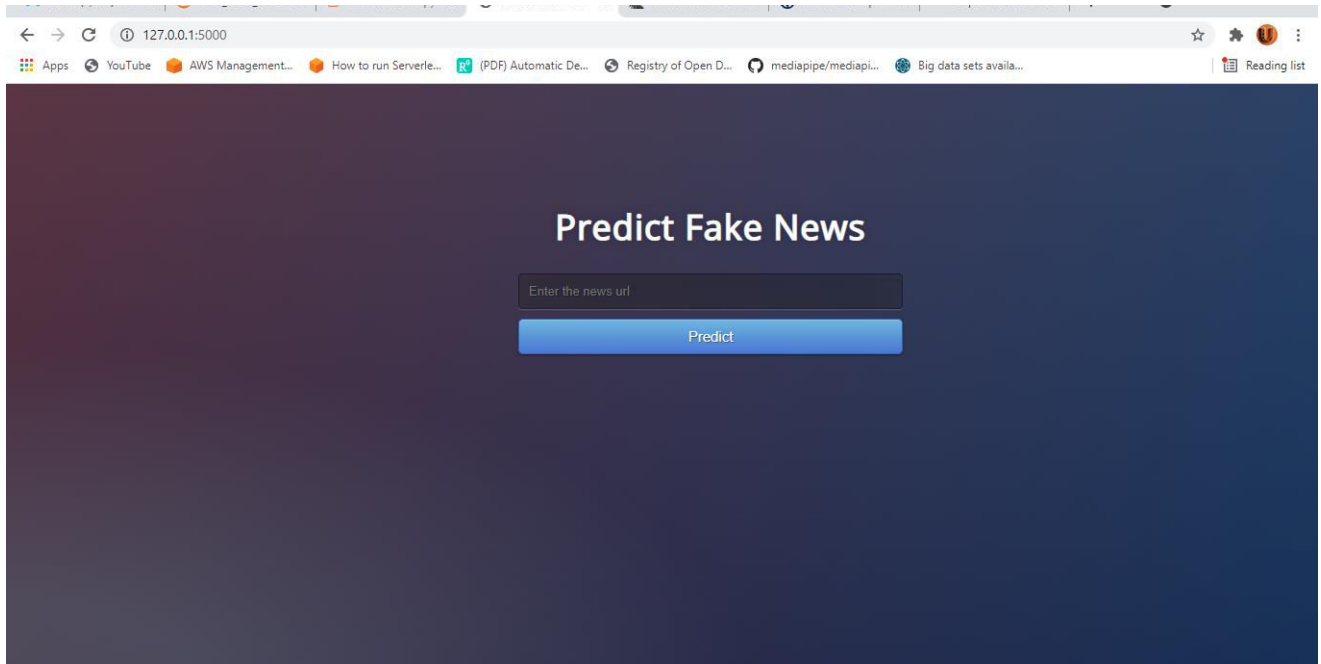


Fig 16. UI of project

We now check few news articles whether they are real or fake.

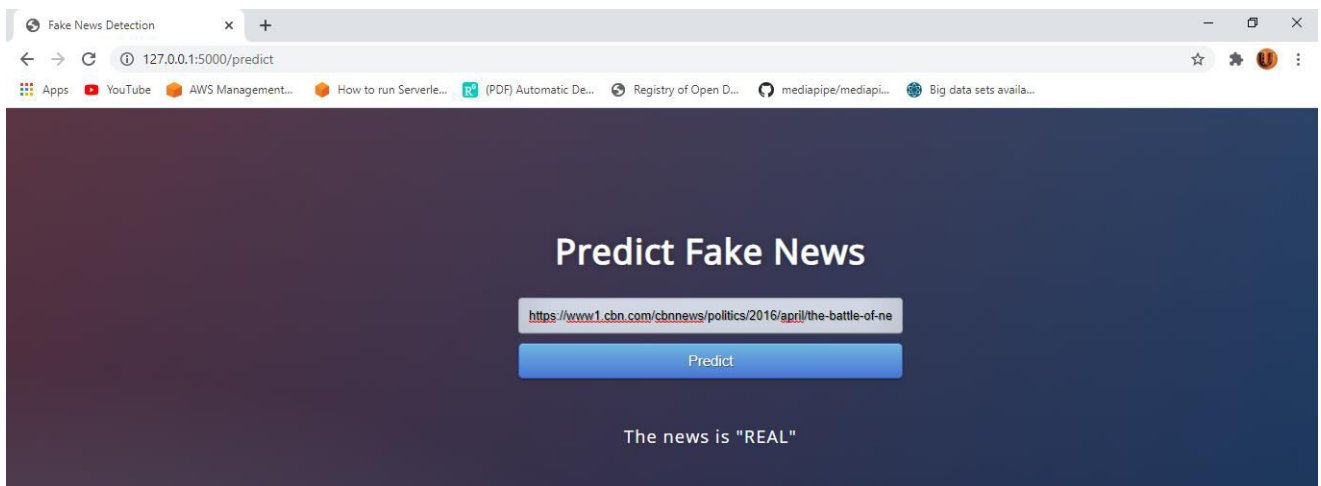


Fig 17. Predicted news is real

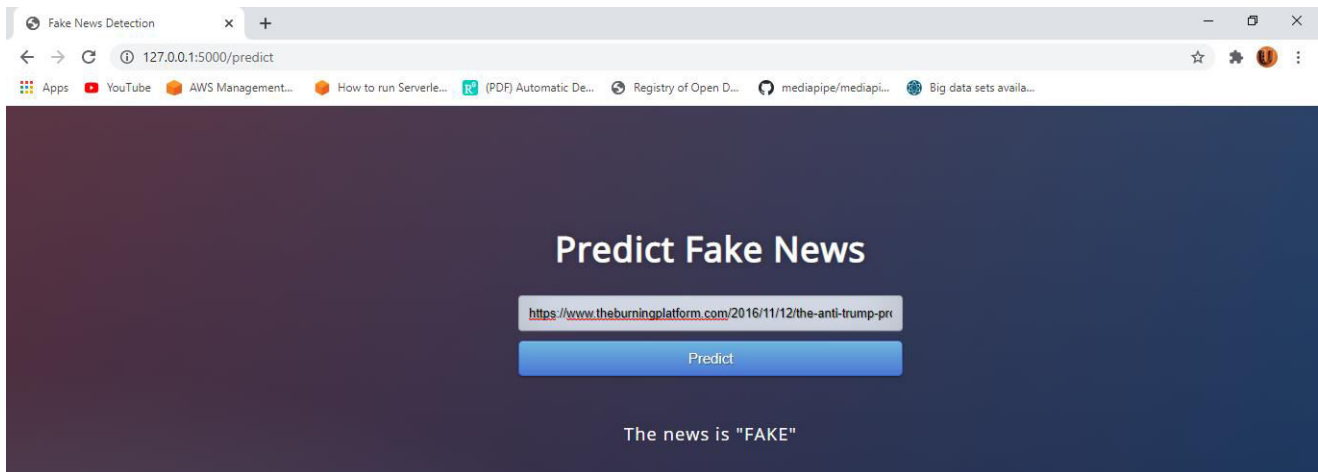


Fig 18. Predicted news is fake

V. CONCLUSION

Within the course of our study, we realized that people have blind trust on posts revealed on social media and therefore it is very important to eliminate the matter of fake news from social media. We presented an outline of different models and algorithms used for identification of fake news on social media. Naive Bayes is suitable for finding multi-class prediction issues and also applicable for Text classification, Sentiment Analysis. Also Logistic Regression is easier to implement, interpret and very efficient to train, therefore by exploitation on top of algorithms we will conclude that any news from a large or small dataset can be classified as fake or real news which in turn helps the users to believe in a particular news.

REFERENCES

1. Janze, Christian and Marten Risius, "Automatic Detection of Fake News on Social Media Platforms", Pacific Asia Conference on Information Systems(PACIS),2017.
2. Ponnurangam Kumaraguru, Aditi Gupta, Henmank Lamda, Anupam Joshi, "Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy", International conference on World Wide Web companion,2013.
3. Aditi Gupta, Ponnurangam Kumaraguru, "Credibility ranking of tweets during high impact events", PSOSM, 2012.
4. Shubham Gupta, Dishant Grover, Anushka Gupta, Deepanshu Kapoor, and Narina Thakur, "A Survey on Identification of Fake News", International Journal for Research in Applied Science & Engineering Technology (IJRASET), ISSN: 2321-9653, Volume-6 Issue-4, 2018.
5. Vasu Agarwal, H.Parveen Sultana, Srijan Malhotra, Amitrajit Sarkar, "Analysis Of Classifiers For Fake News Detection", International Conference On Recent Trends In Advanced Computing(ICRTAC),2019.
6. Subhadra Gurav, Swati Sase, Supriya Shinde, Prachi Wabale, Sumit Hirve, "Survey on Automated System for Fake News Detection using NLP & Machine Learning Approach", International Research Journal of Engineering and Technology (IRJET), ISSN: 2395-0056, Volume-6 Issue-1,2019
7. Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D. Deva Hema, "Fake News Detection using Machine Learning and Natural Language Processing", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-7 Issue-6,2019
8. Faraz Ahmad, Lokeshwar R., "A Comparison of Machine Learning Algorithms in Fake News Detection", International Journal on Emerging Technologies, 177-183,2019.
9. Vivek Singh, Rupanjal dasgupta, Darshan Sonagra, Karthik Raman, "Automated Fake News Detection using Linguistic Analysis and Machine Learning".
10. Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre and Rada Mihalcea "Automatic Detection of Fake News",2018.
11. Monti, F., Frasca, F., Eynard, D., "Fake News Detection on Social Media using Geometric Deep Learning",2019.



INNO  **SPACE**
SJIF Scientific Journal Impact Factor

**Impact Factor:
7.512**

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 **9940 572 462**  **6381 907 438**  **ijirset@gmail.com**



www.ijirset.com

Scan to save the contact details